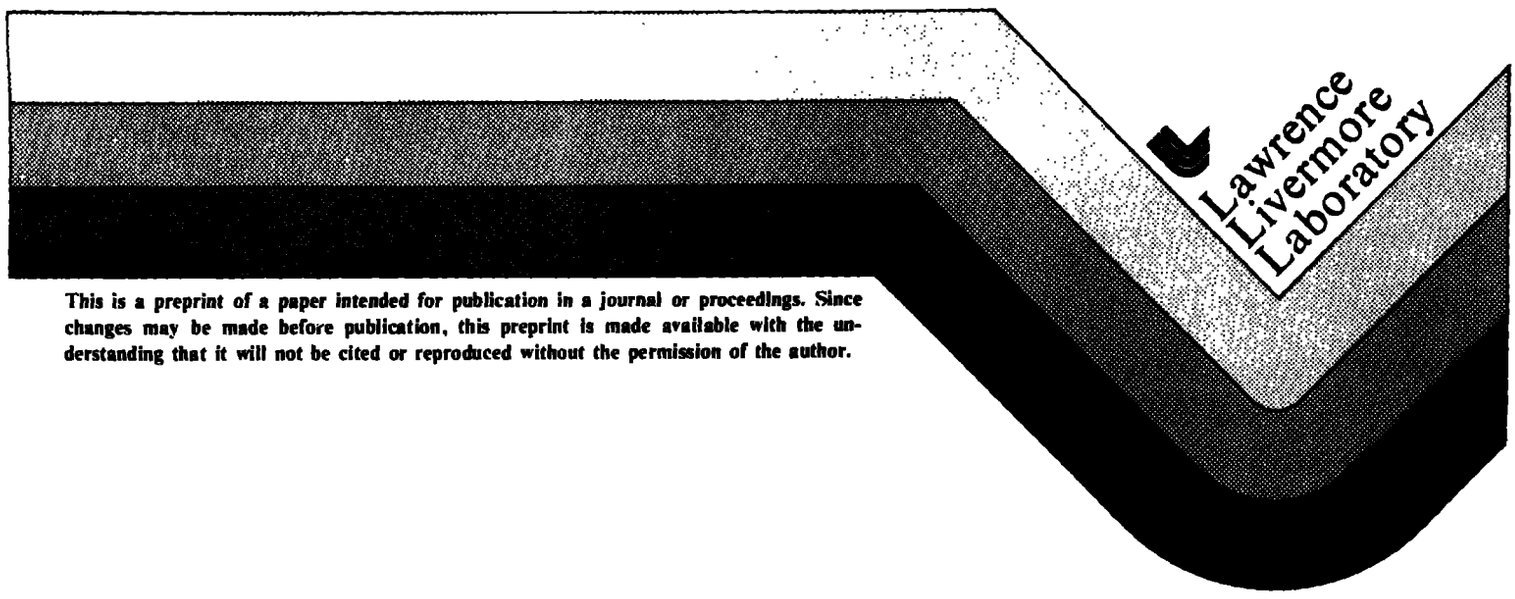
UCRL- 86581
PREPRINT

CONCERNING TECHNICAL MEANS FOR DEALING WITH ASPECTS OF THE NEAR-TERM INFORMATION ONSLAUGHT

Lowell Wood

This paper was prepared for presentation at:
The Conference on Science and the Information Onslaught
Los Alamos National Laboratory
1-5 June 1981



CONCERNING TECHNICAL MEANS FOR DEALING WITH ASPECTS OF THE NEAR-TERM INFORMATION ONSLAUGHT*

Lowell Wood
University of California Lawrence Livermore National Laboratory
Livermore, California 94550

I bring fraternal greetings from the staff of the Lawrence Livermore National Laboratory where, as in all other Laboratories of the University of California, untrammeled intellectual inquiry thrives, except on certain rare and regrettable occasions.

*A paper prepared for presentation at *The Conference on Science and the Information Onslaught*, Los Alamos National Laboratory, Los Alamos, New Mexico, 1-5 June 1981. Work performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under contract number W-7405-ENG-48.

From Where Is The Information Onslaughting?

When considering near-term means for coping with the information onslaught, a prudent student will inquire a bit beforehand as to its nature; when commencing to do so, one is drawn naturally to examine the dominant features of the beast. Guided as always by Lord Kelvin's admonition that the distinction between philosophy and science is that the latter intellectual endeavor has numbers associated with it, one starts inquiring for the quantitative aspects of "the charge of the bits."

If one judges by size scales alone, the information onslaught is perhaps to be identified most aptly with the arbitrarily large amounts of information available from integrating the differential equations describing physical reality to generate histories of particular physical situations characterized by their initial and boundary conditions. It is only little realized even today that the advent of digital computing technology has elevated physics in less than a single human lifetime from natural philosophy to an ability to predict the future of patches of physical reality with degrees of fidelity, detail and omniscence completely beyond the comprehension of the Delphic oracle. To the most profoundly humbling question posed to Job three millenia ago, the human race can now answer 'Yes!', due in large part to the elaboration of physical theory which modern digital computing has made possible. The huge intellectual advances of the past few centuries have brought us knowledge of 'the ordinances of heaven,' and digital computing has taught us much of 'the dominion thereof in the earth' (Job, Chapter 38).

Corresponding to this capability and the value which at least some national governments have attached to it, many of the largest computer systems in the world have been devoted to such scientific soothsaying for the past third of a century. Indeed, it is worth recalling that electronic digital computers were born during the largest scale human conflict in history out of the hope that their possession would confer major advantage, if not victory, and that their youth was generously nurtured for similar reasons when the conflict shifted to less violent forms. At the very least, such a recollection may be well-received at this Laboratory, where modern digital computing undeniably came into its own under the leadership of pioneers such as Nick Metropolis and his colleagues.

The suggestion that physical modelling activity could be at the center of the information explosion gains plausibility when it is considered that there are hundreds of thousands of practicing senior scientists and technologists alive, many of whom are potentially curious as to how a physical situation of interest to him or her will behave under varying circumstances, but who lacks at least some of the resources required to create and study the real situation as desired. Each is thus a potential customer for a super-scale computer system for digital modelling purposes, probably one comparable to the most powerful currently in existence. Moreover, as modelling completeness and integration accuracy can only be finite on any finite speed

computing system, human proclivities toward perfection may guarantee an open-ended demand for such information processing and manipulating capability.

As a specific example, the scientists at my Laboratory at Livermore were provided with digital computing capacity which doubled in effective capacity an average of every 16 months over a two-decade period during the Laboratory's youth, and found this situation very productive; indeed, the cessation of this exponential growth in capability during the past decade was quite traumatic. Moreover, the Livermore experience was not atypical, either in its exponential growth curve during the '50s and '60s, or its relative stagnation during much of the '70s. Sustained exponentiation over many decades certainly is one of the central characteristics of an explosion — perhaps this explosion in information processing capacity for physical modelling purposes is the center of the information onslaught. It's as close as I have been able to locate, and the rest of my talk will take it as a point-of-departure.

All Dressed Up And No Idea Of Where To Go

The currently most powerful computer system for physical modelling purposes is the CRAY-1, which has an incremental production cost of about $3 M. (I remind you that a CRAY-1 does high precision arithmetic at peak sustainable rates comparable to that which could be attained by the entire human race together, each using one of the best commercially available hand calculators; on the other hand, performing high precision arithmetic, even with powerful aids, is certainly not what the human brain does best.) A moderately fortunate and serious physical modelling researcher presently gets use of about 3 percent of a CRAY-1 on the time average; that is, access to digital modelling tools is presently allocated so that prorated incremental production costs of modelling hardware are comparable to the modeller's annual burdened manpower cost to his organization, which is not retrospectively surprising. At the Livermore Lab, those of us in the S-1 Project are presently building a set of computers, each one of which is comparable in physical modelling power to a CRAY-1 but whose incremental production costs are an order-of-magnitude lower, due to the last half-decade's technological advances. This implies that an order-of-magnitude increase in computing power will eventually become available to the average researcher, which will serve to whet appetites — if not completely satisfy them — for 3-dimensional physical modelling at space-and-time resolutions comparable to present 2-dimensional work.

We expect the S-1 work to culminate in the mid-80s with the realization of a computer system with an order-of-magnitude greater computing power than that of a CRAY-1, realized on a single crystal of silicon less than a dozen centimeters in diameter. It is interesting to note that *all* of the technology necessary to realize such a system is currently available, with just one exception: we do not know

exactly how to employ the exceedingly powerful tools which are presently available for patterning silicon.

A few technological considerations may serve to illustrate this tantalizing problem. Present state-of-the-art computer systems contain anywhere from several million to a few dozen million transistors in their central processing units, corresponding to several hundred thousand logic gates or switches and as much as several million bits of Central Processing Unit internal 'state' or specification-of-processing-condition. Each of these gates typically switch information after an intrinsic delay of most of a nanosecond, permitting the CPUs which they constitute to perform an operation in one to two dozen nanoseconds — the characteristic couple dozen gate delays per major computer operation. Now virtually everyone has heard of quantum interferometric ('Josephson junction') digital logic gates operated at liquid helium temperatures which can switch on 25 picosecond time scales; less widely hearalded are room temperature silicon gates which have been demonstrated to operate with 40 picosecond gate delays, moreover implemented in *the* standard commercial mass-production VLSI technology, n-channel MOS. To completely process a full 10 centimeter diameter wafer in this latter type of integrated transistor technology involves a present cost of about $50; a wafer of this size could carry a few of the mid-80s S-1 computer systems to which I have referred. The implications of integrated circuit physical implementation capabilities are utterly manifest in the Figure: continuing of the past decade's merely exponential advance in time of circuit density (and speed, and cost-effectiveness, and energy efficiency, and ...) is the *least* that can be expected during the '80s. It is therefore difficult to see from where the production cost of a state-of-the-art superprocessor will come, a half-decade hence; the cost a decade from now is completely unclear from the present vantage point, a remarkable state of affairs characteristic of revolutionary circumstances.

But why the half-decade delay? If we can even now paint frescos on silicon which not merely delight the eye but emulate major aspects of the brain, what is holding up the advance of this most modern — and most significant — art?

The basic problem is that we have moved all too rapidly from an era of silicon famine — that is, the period in technological history characterized by acute, cost-driven scarcity of logic and memory in computers — to the time of silicon plenty, in which the large-scale integration revolution has already driven the incremental mass production cost of the highest speed transistors below $\$10^{-4}$ each. While we have long known in general what to do when finally entering this promised land, in characteristically human fashion we really did not expect it to happen quite so soon, and we certainly did not adequately prepare for it.

Birth Defects: The Chock in Front of the Wheel of Progress

In retrospect, the genesis of this problem occurred when the advent of the earliest computers successfully drove the improvement of vacuum tubes to performance and reliability levels sufficiently high that computers containing thousands of them could operate for a few hours between comprehensive maintenance periods. This unfortunate — and quite unexpected — technological advance allowed the early computer designers to circumvent the otherwise compelling necessity analyzed by von Neumann and his collaborators to create reliable computing systems out of unreliable components, through the intelligent use of redundancy. Thus, in order to save something of the order of a factor of two in computer construction costs during the past third of a century, we have foregone having highly reliable, single-point failure-immune computer systems during this same period. Though clearly tolerable for this entire period, single-point failures are currently becoming *the* impediment to further technological advance. Now that the incremental construction cost of a state-of-the-art computer is approaching the cost of a single man-day of maintenance technician time, the economic imperative has flipped direction and presently points the same way as the esthetic one always has: towards single-point failure-immune computers.

There is an extremely important practical consequence of single-point failure-immunity of an operating computer associated with the extension of this feature all the way back to when the brand-new computer is first turned on. Recall that most modern microprocessors — the ancestors of supercomputers-on-a-chip — are born dead; due to the presence of one or more fabrication flaws of an irreparable nature, they are inoperable in some significant sense. This is well-known to semiconductor specialists as "the yield problem". Through the present time, the power of microprocessors is limited by how much circuitry can be integrated together before the likelihood of a fatal fabrication flaw appearing at random within its implementation becomes too close to unity, i.e., the 'yield' of fully functional chips from a wafer of them becomes too low. However, a single-point failure-immune supercomputer-on-a-chip may have such congenital defects, but they will be masked — made effectively invisible from the outside of the chip — by single-point failure-immune design of the chip's circuitry.

Indeed, this immunity to a single birth defect is already more the rule than the exception in the most modern semiconductor memory chips (64 K static RAMS). Containing up to a few hundred thousand high speed transistors per chip, typically 97 percent of these chips are stillborn, with at least one fatal manufacturing flaw somewhere on them. With the built-in, highly automatically implemented ability to substitute a new set for a single memory cell's half-dozen transistors, 80 percent of these dead-at-birth chips may be revived to lead useful lives of full duration and capability. Interestingly, the increment to the transistor population of a memory

chip to raise live chip yields from 3 to 80 percent is typically less than 2 percent — the intelligently applied redundancy tax is really very low!

The huge gain in cost-effectiveness associated with such minimal redundancy is indeed striking; on the other hand, it is intrinsically limited to computer memories, and CPU logic may have to be duplicated or triplicated in order to become comparably robust against all first contingencies. That is the bad news — the good news is that over 90 percent of the transistors in a modern CPU are actually used to constitute memory elements, with only ten percent or less devoted to making logic gates: modern computing structures are really evolving rapidly towards intelligent memories, with large bricks of memory being mortared into computing structures with small dabs of logic. The pervasive use of redundancy to safeguard against the deleterious effects of all first — and quite possibly even multiple — contingencies in computers to be created during the next half-decade is thus likely to increase their total transistor populations by a few tens of percent, and certainly by less than two-fold. Since the number of transistors which can be created on a given area of silicon doubles in little more than a year these days, the 'premium' in transistor population required to realize truly failure-immune supercomputers-on-a-chip will be readily affordable, particularly since the alternative is manufacturing yields of a tiny fraction of one percent and frequent, irreparable failures-in-operation of the miniscule fraction which are born functional.

The work of the next few years in realizing supercomputers-on-a-chip will thus be focussed primarily on creating highly automated means of introducing the required degrees of redundancy into supercomputer designs, and secondarily on projecting the resulting designs onto monolithic, nearly two-dimensional semiconductor-based surfaces. The S-1 Project, as well as similar undertakings elsewhere, sees these two tasks as *the* challenge to computer technology in the early '80s. As I remarked before, realizing such designs in extremely high performance silicon can be accomplished with technological means already in hand, e.g., discretionary e-beam direct-write-on-wafer systems generating fractional micron effective channel length *n*MOS transistor patterns, interconnected in the usual fashions.

Appropriate Behavior Upon Entering the Promised Land: MIMD Networking

So here we are a half-decade hence with the ability to crank supercomputers-on-a-chip out of our silicon foundries at least as agilely as a short-order cook whips up a stack of pancakes. Every serious professional has ten CRAY-1-equivalents of personal computing power in his or her office later in the decade, at a cost comparable to that of current high-performance hand calculators. Is this the extent of the potential of the ULSI revolution, the full implication of attaining the Ultra-Large Scale Integration plateau? Is there no practical way to increase effective

computing power faster than the underlying logic technology improves?

The S-1 Project, along with a growing band of others elsewhere — and I was delighted to hear yesterday that Jack Schwartz is among them — believe that confederations — federal unions, if you will — of N nominally independent processors can be designed, built and programmed to support reasonably close to N-fold speed-ups in the wall clock rate at which most problems of interest can be solved.

Very straightforward considerations of time-average computing efficiency of a typical processor suggest — as Jack sketched yesterday — that one should have as much computing power in each single member of the multiprocessor federation as can be provided in a reasonably cost-effective fashion, and so the first S-1 multiprocessor is being implemented during the coming year with processors of roughly the scientific computing rate of a CRAY-1.

Considerations of proper balancing of hardware and software costs suggest that an experimental multiprocessor system should be of sufficiently great performance scale to justify the costs of developing serious systems and applications software for it, and so the first S-1 multiprocessor is being constituted of 16 processors.

Similar considerations suggest the appropriateness of as high a quality interconection between processors and memory moules as is economically feasible, and so the highest connectivity switching network — a full crossbar switch — is being provided to simultaneously interconnect all processors in any one-to-one-onto mapping to all memory modules in the S-1 multiprocessor, to arbitrate fairly in the event of momentary request conflicts, and to provide similar interconnectivity of processors among themselves.

It is worth noting in this latter respect that full crossbar switches are an eminently practical means for connecting together even quite large federations of processors and memories. While those who study data switching networks professionally find crossbars too simple to provide much grist for their mills, and thus disdain them in favor of more exotic interconections, technology quite perversely makes crossbars too cheap to forsake for more cleverly designed but intrinsically lower performance interconnection approaches.

As a specific example, we find that full crossbar connectivity of our S-1 16 member multiprocessor has a quadratically growing fraction whose cost is somewhat less than one percent of that of the multiprocessor system as a whole; in other words, full crossbar connectivity could be provided with our current S-1 technology to interconnect 10^3 processors with each other and with their memory modules before the cost of the switching network would rise to equal that of the processors and memory modules. Only when providing for networks of more than of order 10^3 processors with 1981 technology would one need to consider less costly approaches, such as the Perfect Shuffle idea which Jack sketched yesterday. Alternatively, one

might choose to interconnect a number of networks, each constituted of several dozens to a few hundred crossbar-connected processors, into super-scale networks, again using crossbars to net the networks together.

But granted that one can build and interconnect such huge piles of hardware. Can such interconnections possibly be programmed with finite effort, can they be exercised with reasonable efficiency, and can their enormous number of possible failure points be coped with? Interestingly enough, the answer appears to be 'Yes' to all three of these questions.

Programming for the S-1 type of Multiple-Instruction-stream/Multiple-Data-stream (MIMD) multiprocessor turns out to be quite straightforward: a small number of grammatical constructs of a comment nature appended to the standard grammar of any high-level algorithmic language — such as FORTRAN or Pascal — suffices to extend it to an efficient multiprocessor programming language. Programs in this language may be readily compiled for execution on a multiprocessor with a processor population which is dynamically variable from 1 to N *during* program execution. The increase in programming effort to fomulate problems for a multiprocessor, relative to that for a single processor, is still being investigated, but appears to be of the order of 10 - 25 percent for typical scientific algorithm mixes.

Multiprocessor utilization efficiency studies also turn up preliminary but quite optimistic results. Examinations of both analytic and simulation natures indicate that a reasonably wide variety of scientific algorithms can be executed on a 16 member MIMD multiprocessor of the S-1 variety with average hardware utilization efficiencies of between 55 and 95 percent. The overall time-averaged efficiency of MIMD multiprocessor execution of a Livermore Lab benchmark two-dimensional Lagrangian shock hydrodynamics and ADI heat transport physical simulation code was found to be 75 percent, i.e., it ran 12 times faster on a simulated 16 member S-1 multiprocessor as it would have executed on a single member processor. Jack quoted closely comparable results to you yesterday, and still other researchers find similarly. Clearly this is not a universally applicable number, but it is representative of mainstream scientific computing, and as such is very hopeful.

The Gordian knot of the multiprocessor reliability question can always be sliced apart by checkpointing a calculation to redundantly reliable memory adequately often, and then exploiting the variable processor population feature of the multiprocessor's problem partitioning software to always use all of the multiprocessor's healthy processors. More elegant approaches — akin to untying the Gordian knot — which are too complicated to sketch in this short talk also appear to be practical.

In summary, the MIMD type of multiprocessor appears to open up a track of a highly practical nature leading to comparatively huge increases in effective computing power, even though the performance of the underlying logic may improve only relatively slowly. As such, it is *the* method of choice into the foreseeable

future for translating the continuing enormous gains in digital logic density and cost-effectiveness into correspondingly great advances in useful computing power. I therefore suggest that it will be at the center of means of dealing with near-term aspects of the information onslaught.

Meanwhile, Back on the Hardware Frontier ...

However, likely advances in the underlying hardware technologies should not be ignored. Digital logic gate speeds seem likely to advance by at least an order of magnitude from the 25-40 psec rates at which the fastest JJ, GaAs and Si gates work at present, though exploitation of the one psec JJ switching rates very recently reported, by the development of 'nearly short' channel GaAs MESFETs comparable in scale to the currently most advanced Si gates, and via continuing advances in fractional micron Si lithography. Advances to sub-picosecond gate delays via optical digital logic, though of higher technical risk, nonetheless seem likely in the '80s through work underway at Livermore and elsewhere.

It is also crucial to note that the energy needed to perform an elementary logic operation, presently a few femtojoules, continues to decrease rapidly with technological advance. There is still a gap of a million-fold between the energy dissipation of the best digital logic gate technology and that theoretically required. Closing most of this six order-of-magnitude gap with molecular-scale transistor elements which are feasible to fabricate and interconnect in exceedingly large quantities is a surpassingly exciting prospect, one being pursued at Livermore and probably elsewhere.

It is easy to forget that memory is part-and-parcel of computing, and that a powerful CPU with insufficient memory attached to it is like a weight-lifter with paralyzed legs: a few feats can be performed impressively anyway, but in most performance is degraded, and some things are just impossible. Indeed, one of the cardinal empirical relations of modern digital computing, Amdahl's Rule, states that a computer must have approximately one word of high speed memory attached to it for every instruction per second of computing which it is expected to efficiently perform. It is just because memory is so regular, so symmetric in nature, that it is easy to design and efficient to manufacture; memory technology has characteristically paced digital semiconductor technology development as a whole. We thus tend to take it for granted, relative to the "problem child," digital logic, with its highly irregular, quasi-random nature.

High-speed semiconductor memory is presently quite fast — data recall times range from somewhat under 10 to about 100 nanoseconds — and very cheap — about $\$10^{-2} - 10^{-4}$/bit (depending on the required speed of data recall). Moreover, another famous empirical rule due to Noyce, the co-inventor of the integrated circuit,

states that memory price and the silicon area required per bit stored drops by two-fold annually, on the average; this rule has been in effect for the past decade, and continues to be valid through the present.

Interestingly enough, some types of semiconductor memories require no power to retain data (but can only be written perhaps 10^5 times, though read as often as desired), others retain data without power but can be re-written only quite slowly (though read very rapidly and arbitrarily often), while still others are completely high speed read/write units and only require a few nanowatts per bit to retain data indefinitely. These properties are *all* of interest in creating appropriate memory systems for very high performance computer systems for a reason both rather profound and pervasive: computers, like the people who program them, tend to process data in a peculiarly orderly fashion. Data being processed at any particular moment tends to be closely associated — by some memory metric or another — with data just processed and with that about to be processed. Perhaps such time-sequenced processing of mostly 'nearby' information is intrinsic to most all intelligent manipulation of data; somewhat more likely, it's a subtle feature of how human brains store or retrieve data.

In any event, computing almost exclusively on 'nearby' data is the way it's observed to take place in nearly all applications, and it permits the extraction of 'locality' from data masses and the exploitation of such locality via hierarchies of memory having differing costs, power requirements and recall time features. Each lower layer of such memory hierarchies — starting from the top one serving immediate CPU needs — has exponentially greater storage capacity *and* greater required time-to-recall than the layer just above it. The effect of a very capacious and inexpensive memory with only minimal penalties in *average* recall time for a datum are thereby gained: typically 99 percent of all data needed by the CPU are found in the fastest memory, and typically a 10-fold penalty in retrieval time for one percent of the data requests (or a 1000-fold penalty for 0.01 percent of the data requests) imposes only a 10 percent average latency penalty while buying increased memory capacity at a 100-fold lower per-bit cost, in moving from one layer in the hierarchy to the next lower one. Very many types of memory technology thus find natural niches in this hierarchy, and enrich its potentialities in the process.

But how much memory capacity of what speed and unpowered data retention characteristics will the computer systems of the '80s need, and how much is technologically accessible? Amdahl's Rule provides a time-tested lower bound, and John McCarthy has suggested a reasonable upper bound: a computer should be able to recall its entire previous state from the time it was first put into operation, akin to the most legendary human recall capabilities. Since a 1981 state-of-the-art computer changes an instantaneous state size of the order of 10^2 bits about 10^8 times per second over a really useful life of the order of 10^8 seconds, this would require a memory capacity of the order of 10^{18} bits.

I will state without proof for reasons of brevity that it is possible to construct with existing technology a photochemically based, molecularly encoded memory system having an order-of-magnitude larger capacity — an Einstein of bits — which could be retrieved on sub-microsecond time scales and whose active medium would be a fractional cubic meter in volume. If you are more conservative and inclined to just write a purchase order for a more modest system, a very large and most reputable high technology company is prepared to deliver a functional prototype of a 10^{15} bit capacity, photochemically-based storage system for something of the order of $30 M by mid-decade with potentially a microsecond recall time, as at least some of you are already aware. By way of comparison, I will remind you that a human who could perfectly recall absolutely everything which he or she sensed during a typical lifetime would have stored something of the order of 10^{14} bits; most of us recall something of the order of 10^{7} bits on 1-10 second time scales with high error rates, and, even more raggedly, perhaps 10^{10} bits total, over 10-10^{5} second recall periods.

I would say more on this subject, but Edward will be treating it exhaustively in his talk on Thursday evening, and prudence — if not courtesy — requires that I forebear.

Now That You Have Everything You Ever Wanted ...

So we can realize at mid-decade computing systems on a single chip which are an order-of-magnitude more powerful than the best ones at present for an incremental cost not greatly in excess of $10. So we can interconnect many hundreds of them intimately and for larger numbers somewhat less closely. So what? Everyone then has 100 million Floating Point Operations Per Second — 100 megaFLOPS — of computing power in his/her office terminal, whose cost, mass and power consumption are probably dominated by the visual display unit. Is this where the leading edge of the information onslaught will be in the second half of this decade?

I suggest that, in terms of sheer megaFLOPS and Millions of Instructions Per Second, the answer is almost certainly 'Yes,' just as microprocessors, rather than CRAY-1s, represent the bulk of planetary computing capability at present. However, the seeds of the future will be sprouting elsewhere, with one such seedbed being the use of advanced networking techniques — in both hardware and software — to effectively organize large numbers of superprocessors-on-a-chip into single computing systems running single problems of sizes which are titanic, even by present scales.

A quarter century of observing and dabbling in the American computing scene is the basis for my remark that it is easy to get the Federal Government to put up $\$10^{6}$ for a single computing system, challenging to raise $\$10^{7}$ for such purposes, and

effectively impossible to get Uncle Sam to spring for 10^8 for any computer system; no other patron is in the game, as a practical matter. Incidentally, this little lemma has remained valid while the dollar has changed in real value by a half-order-of-magnitude, and has a corollary due to Seymour Cray, usually quoted as "A top-of-the-line computer system has always cost $10 million, and always will." When the day arrives several years hence that a CRAY-1-equivalent of computing power costs of the order of $1, what can be done with the million CRAY-1 equivalents of computing power that a relatively easy-to-get megabuck will buy?

I know what *I* will do with it, and I will sketch this briefly because I suspect that my tastes in physical modelling are not totally atypical. With two of the six orders of magnitude of computing power gained over the best present systems, I will buy adequate spatial resolution in the presently lacking third dimension. With two additional orders of magnitude, I will purchase a factor-of-three average improvement in space and time resolution in these four dimensions, relative to that enjoyed on present 2-D problems. With the fifth order-of-magnitude, I will get my simulation problems run to completion in a few tens of minutes rather than the several hours which they tend to require on a CRAY-1. With the final order-of-magnitude, I will buy a corresponding increase in the richness of the physical phenomena which can be modelled at a given logical point mesh over a time integration interval, thereby permitting *really* quite extensive population kinetics, or Schrodinger equation solving, or semiconductor device simulation, or whatever, to be performed. Note that such an allocation of aggregate processing power is compatable with having 10^5 interconnected processors at work — by hypothesis, there is something of the order of 3×10^7 logical mesh points in problems of interest, so that each of the 10^5 processors computes for an efficiently large number of grid points.

Others will choose to allocate the capabilities of such a $10^7/10^8$ megaFLOP system (depending on whether it is doing scalar or vector processing) quite differently, and it is clear that their freedom to efficiently do so won't be abridged by the MIMD multiprocessor structure, at least within very wide limits.

But what then? How will additional gains in multiprocessor hardware and software organization be put to reasonable use, to say nothing of the two additional orders of magnitude gain due to improvements in logic speed which can be foreseen during the '80s with fair confidence? I at least do not have particularly inspired suggestions. How to use the upcoming three-to-six orders of magnitude in effective computing power seems clear. How to reasonably employ a billion- to a trillion-fold more computing power — or memory capacity — than is currently available exceeds my imagination at present. Ask me again in a decade, please.

But Is There Life After 1991?

I would now like to briefly invite your attention to some of the implications of near-term advances in human ability to deal with the information onslaught. In particular, consider the magnitudes of the switching rates and memory capacities which we have just been considering relative to those of the human brain. The brain — and specifically its cerebrum — contains something of the order of 10^{12} neurons, each of which can switch at most about ten times per second; each of us can thus perform something of the order of 10^{13} elementary switching operations per second, a slightly awesome data processing rate, even for noisy logic elements. Assume now that each neuron makes switching decisions based on 10^2 bits of internal state — specifications as to how much rest it got the night before, the local glucose concentration and oxygen tension, how recently it fired, the present and recent state of excitation of its dendritic tree, how much and what particular flavor of memory protein, RNA or whatever it contains, etc. — recent research seemingly suggests that a neuron's state may be represented by more like 10 than 100 bits, but let's optimistically assume a relatively rich internal state. Note in passing that the ratio of memory bits to logic gates in the brain is the same order-of-magnitude as in modern computers CPUs, which I characterized earlier as ever more like intelligent memories — the bit-to-gate ratios of both information processing systems are of the order of a few dozen.

We have been considering superprocessors-on-a-chip whose 1981-technology gates (and associated interconnections) are about a factor of ten smaller in linear dimension than neurons, and which switch at least 10^9 times faster while consuming about 10^7 times as much power. A $\$10^6$ cost, late-1980s MIMD multiprocessor composed of such elements would contain 10^5 monolithically integrated superprocessors, each having 10^6 such switches and 10^7 bits of internal state; this is about 10 percent of the upper-bound estimates which I just cited of the logic switch and internal state bit counts of the human brain. If it is possible to allocate the decision-making work of 10 neurons to an average of a single silicon gate — just as we presently assign many logical mesh points to a single CPU in MIMD processing of physical models — and if 10^7 additional, non-state-specifying bits of memory are appended to each CPU chip, then this MIMD multiprocessor system might reasonably be expected to compute at least 10^6 times faster than a human brain: its ten-fold fewer switches each operate 10^9 times faster than do a neuron, and 100 gate operations conditioned by 3000 bits of total state is surely a very generous estimate of the complexity of a tenth-second of data processing operation of a single cerebral cortex neuron. Moreover, the (non-state) memory capacity of the total system would be adequate to store *all* the sensory information gathered by a human over an entire lifetime. Built in 1981 technology, this MIMD multiprocessor system would occupy about 10 cubic feet (including provisions for cooling), and would require about a megawatt of electric power. As the trend line in the Figure suggests, at least a thousand-fold

reduction in these numbers by 1991 is to be expected.

I really cannot suggest to you what a million-fold increase in the computing rate of a human brain-equivalent would portend; I have only a hazy intuition as to what a ten-fold increase in human information processing rate would imply. Indeed, I have only been able to find one quantitative hint in this respect in the published literature, which rather cryptically suggests that God perceives duration — e.g, thinks — only a few hundred thousand times faster than do people (Moses, Psalm 90). It gives one pause for more than a little thought that $\$10^6$ of computing hardware a decade hence may be reasonably expected to bring us to a divine level of information processing capability.

But you might respond that we have no detailed idea of how the brain works, and thus are completely unprepared to provide the software to drive the hardware just scoped out; you might even protest that we do not presently have more than a sketchy, highly incomplete wiring diagram for the human brain's cerebral cortex. My present answer would be that the information content of the entire human genome is only 10^8 bits, and it is difficult to believe a priori that much more than 10^7 bits — 10 percent of the entire set of human genetic information — are devoted to specifying the structure of cortical neurons, their various programs, *and* their interconnections; after all, instructing the liver's mitochrondria how to tranform fine liquor into a hangover, telling toes how to grow toenails and the cerebellum and the spinal cord how to potentially carry out the real-time control of a ballet dancer all are competing for their share of the 10^8 bits. I would then add that a large physical modelling program of 50,000 lines of FORTRAN expands to a few hundred kilowords of program storage, or about 10^7 bits of reasonably non-redundant information. That a single motivated and intelligent human working for a period of the order of a year can create a structured set of 10^7 bits of globally directed, highly interrelated bits I find to be notable in this context — man-years, not man-centuries, of high quality human effort should suffice to write an engineering specification of the human cerebral cortex's structure and function which would be adequately detailed — and accurate — for emulation programming purposes. Such work may get underway seriously once microneurophysiologists have given us a few more hints, hints that I expect will be forthcoming during a normal course of events in the present decade.

Aren't Those the Opening Bars of the Finale?

I, therefore, am led to the conclusion that two of the highest leverage areas of applied research during the '80s will deal with developing a reasonably high bandwidth, full duplex, time-stable interface between interesting portions of neural nets of the human central nervous system and semiconductor chip 'edges' — bonding

pads or whatever — and with creating a reasonably compact, efficient, implantable glucose-oxygen fuel cell, or its functional equivalent, for supplying something like ten watts of average electrical power production within the human body. This conclusion seems compelling simply because the far edge of near-term dealing with the information onslaught will certainly involve at least some humans carrying substantial silicon implants, *mental* protheses with information processing and data storage capabilities which will greatly outstrip our native ones in certain useful — possibly initially highly specialized — respects.

Note that paying even as much as a few femtojoules per logic switching operation and a nanowatt per bit of memory — i.e., using completely unimproved 1981 technology — potentially buys 10^{10} bits of instantaneous recall and multiple CRAY-1s worth of processing capacity in a package comparable to that of a cardiac pacemaker, moreover, one whose glucose and oxygen requirements for power supply would not be much greater than that of the human brain (a dozen electrical watts would suffice).

A decade hence then, human-silicon hybrids may be walking around, looking just like you and me, but each endowed with a thousand times the short-term memory and a ten billion times the specialized data processing power as each of us enjoy today. There is thus sustantial reason to believe that these hybrids may be more "successful" in some of the Darwinian or sociobiological senses than will ordinary humans, with our information processing and storage capabilities unaugmented by silicon implants.

It would be surprising if it did not quickly become feasible for such hybrids to rather frequently re-program their implants so as to conduct an ever larger fraction of the exercising of their organic brains which we call 'consciousness' in the logic sections of these implants, and to store an ever larger fraction of memory-of-events and the results from conscious processing of it which we call 'self' in the memory portions of these implants. The ever greater speed and fidelity of processing and recall which is likely to result as the technology underlying such implants and their programming continues its exponential advance in time will not only strongly stimulate such a transfer of mental functions from a tissue base to a semiconductor one, but clearly will soon alter the nature of such human-silicon hybrids beyond our present ability to imagine.

Of even more profound consequences from our present world-view, perhaps, is the very real transmigration of the human soul from its present seat in the all-too-mortal flesh to a new home in failure-proofed hardware which will thereby occur — not only will the human ego thereby be able to naturally and unambiguously survive the inevitable dissolution of the body which initially sustains it, but it may choose to transcend the flesh just as quickly post-implant as valued memories are recalled from neuronal storage for purposes of being moved into a silicon base. One aspect of

near-term human dealing with the information onslaught may thus be physical immortality and perfectability of a hitherto little anticipated nature. The concomitant ability to do high bandwidth transmission of the bit pattern representing one's entire consciousness-and-self around the physical universe at lightspeed on modulated electromagnetic radiation from host hardware at one location to that at another, possibly far-distant one, may prove to be of greater practical, if lesser philosophical, importance.

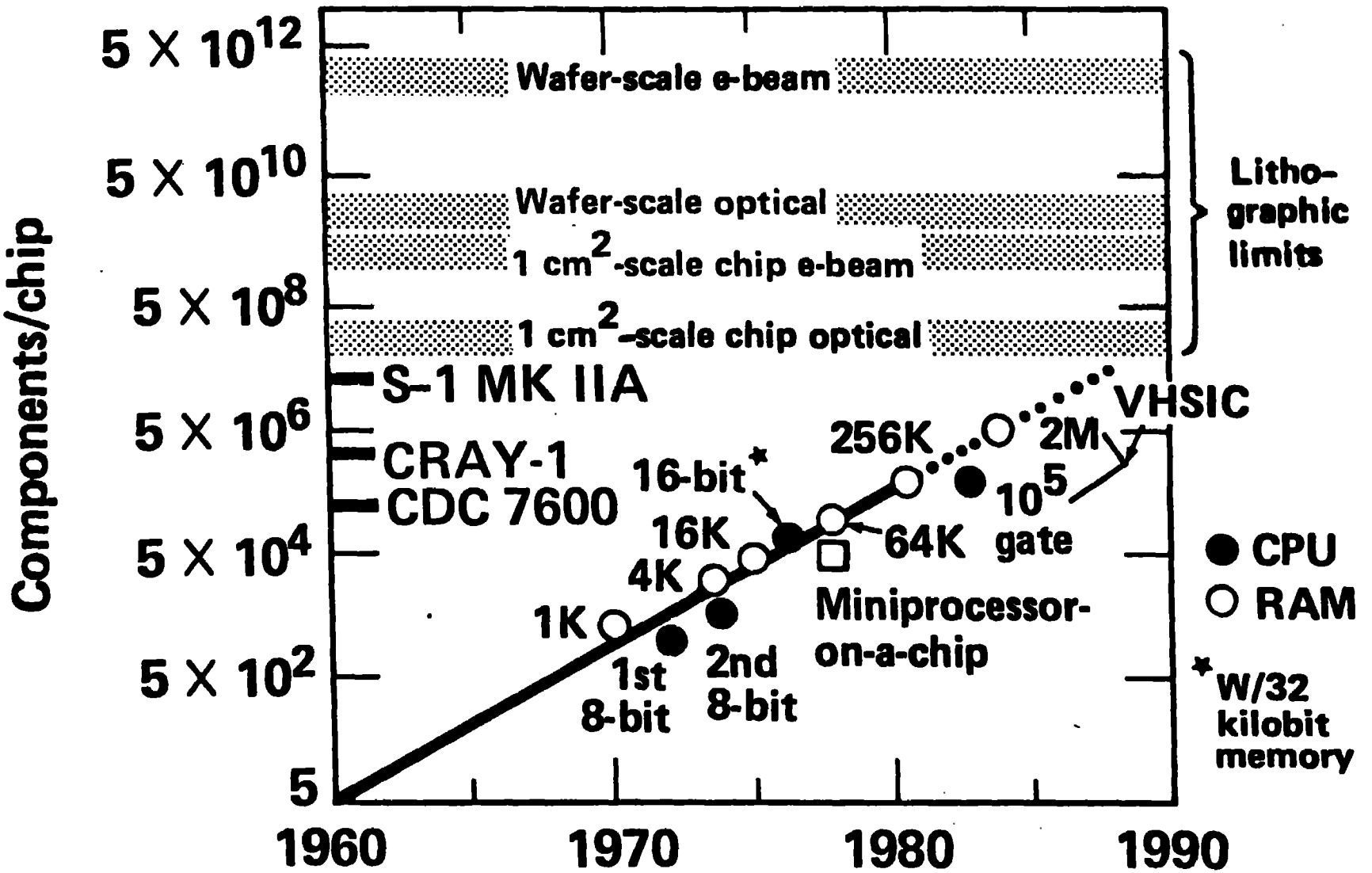
That human transfiguration of such a profound degree might occur by the close of this century on a technological base only relatively little advanced over the existing one I find to be most striking. I therefore suggest to you in conclusion that only the near-term aspects of dealing with the information onslaught are of human interest, and that the next step is about to be taken in the eons-long procession of species: we will be so altered as a race so soon by coping with the information onslaught that taxonomists a century hence will declare *Homo sapiens sapiens* to be extinct. For the first time in the history of life on this planet, though, the extinction of a species may be voluntary, moreover on an individual-by-individual basis and within a very few generations. Neither artificial nor natural intelligence may be found on Earth a century hence, though intelligence of degrees of which we do not presently dream may then grace the home planet of Man.

Why so quickly? Consider the effective time required to usefully twiddle bits via random mutation and natural selection in a strand of primate DNA as compared to that needed to purposefully alter bits in a computer's memory. A new species of computer-enhanced humans — unable to effectively exchange information with earlier versions or unwilling to make the attempt — can be expected to arise literally overnight, not just once, but as frequently as such entities can devise and agree upon new modes for continuing the exponentiation in time of their information processing capabilities. In brief, 'though we may not all die, we will all be changed, and in a moment, in the twinkling of an eye ....'

It was more than a third of a century ago, Edward has recalled, during a wartime lunch break at this Laboratory that Enrico Fermi opened the modern discussion of intelligent life elsewhere in the universe by inquiring in a context so empty that his succinct question served very naturally to define the new one, "Where is everybody?" Extensions of the military work in which he and his colleagues were then engaged have so captured the imagination of the intelligentsia since then that it is almost universally assumed that civilizations elsewhere in the universe inevitably attain the capability of nuclear (or perhaps biological) warfare, thereupon to quickly and irrevocably suicide; thus, we never hear from them. Perhaps the 'standard' evolutionary track of intelligent planetary life is somewhat different, and is to be inferred from the not unrelated premises that intelligence strives to process ever greater amounts of information more and more effectively, and that the rise of nuclear weapons technology in a typical civilization occurs at most a short time

before the advent of electronic digital computing. I therefore suggest to you that a more valid response to Fermi's inquiry may be, "Pity forbids their appearance locally, for yet a little longer."

DIGITAL LOGIC CAPABILITY COMMERCIALLY AVAILABLE ON A SINGLE CHIP



FIGURE